| | |
|---|---|
| **Iniziato** | domenica, 5 febbraio 2023, 14:30 |
| **Stato** | Completato |
| **Terminato** | domenica, 5 febbraio 2023, 15:23 |
| **Tempo impiegato** | 53 min. 36 secondi |
| **Valutazione** | **17,00** su un massimo di 17,00 (**100**%) |

**Domanda 1**

Risposta corretta

Punteggio ottenuto 4,00 su 4,00

Un'agenzia immobiliare desidera indagare quali sono le caratteristiche che attirano maggior interesse nei clienti sul loro sito. Per un campione di 47 immobili vengono raccolte diverse caratteristiche e viene registrata l'informazione di quanti clienti contattano l'agenzia entro 24 ore dall'aggiunta dell'immobile sul sito. In particolare vengono registrate le seguenti informazioni:

- `n_cont`: il numero di clienti che hanno contattato l'agenzia entro 24 ore dall'aggiunta dell'immobile sul sito
- `n_foto`: il numero di foto inserite nel sito
- `m2`: la dimensione dell'immobile (in $m^2$)
- `type`: l'informazione se l'immobile è una casa singola, una casa bifamiliare o un appartamento

La prima analisi si sofferma sull'effetto che ha il numero delle foto nell'annuncio (`n_foto`) sul numero attesi di contatti con l'agenzia:

```
dim(df)
```

```
 [1] 47  4
```

```
fit_foto <- glm(n_cont~n_foto,family = poisson, data = df)
logLik(fit_foto)
```

```
 'log Lik.' -67.488 (df=2)
```

```
deviance(fit_foto)
```

```
 [1] 47.19
```

Si desidera usare i modello `fit_foto` per derivare una stima del predittore lineare del modello e del valore atteso di contatti in agenzia per un immobile con 30 foto:

```
coef(fit_foto)
```

```
 (Intercept)      n_foto
     -1.8593      0.1756
```

```
predL <- predict(fit_foto, newdata = data.frame(n_foto = 30), type = "l")
predE <- predict(fit_foto, newdata = data.frame(n_foto = 30), type = "r")
```

```
predL
```

```
 [1] NA
```

```
predE
```

```
 [1] NA
```

Infine si stima un modello in cui tutte le variabili a disposizione sono inserite nel modello:

```
fit_all <- glm(n_cont~.,family = poisson, data = df)
```

```
deviance(fit_foto)
```

```
 [1] NA
```

Si indichi il valore del BIC per il modello `fit_foto`. Si indichino poi le stime del del predittore lineare del modello e del valore atteso di contatti in agenzia per un immobile con 30 foto. Infine si indichi se la devianza per il modello `fit_all` è maggiore, minore o uguale alla devianza per il modello `fit_foto`.

a. Il valore del criterio BIC per il modelli `fit_foto`: .

142,676

b. La stima del valore del predittore lineare predL: .

3,4087

c. La stima del valore del valore atteso predE: .

30,226

d. ◉ deviance(fit_foto) > deviance(fit_all). ✔

○ deviance(fit_foto) = deviance(fit_all).

○ deviance(fit_foto) < deviance(fit_all).

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: deviance(fit_foto) > deviance(fit_all).

BIC: \[-2*logLik(\mathcal{M}(p)) + log👎  * p\]

Stima per il predittore lineare:

$$-1.8593 + 0.1756 * 30$$

Stima per il numero atteso di contatti:

$$\exp(-1.8593 + 0.1756 * 30)$$

La devianza funge la stessa funzione del RSS nei modelli lineari: più variabili si inseriscono nel modello più diminuisce la devianza.

a. 142.68
b. 3.41
c. 30.20
d. Vero / Falso / Falso

Domanda **2**

Risposta corretta

Punteggio ottenuto 5,00 su 5,00

Un'agenzia immobiliare desidera indagare quali sono le caratteristiche che attirano maggior interesse nei clienti sul loro sito. Per un campione di 37 immobili vengono raccolte diverse caratteristiche e viene registrata l'informazione se un possibile acquirente prende contatti con l'agenzia entro 24 ore dall'aggiunta dell'immobile sul sito. In particolare vengono registrate le seguenti informazioni:

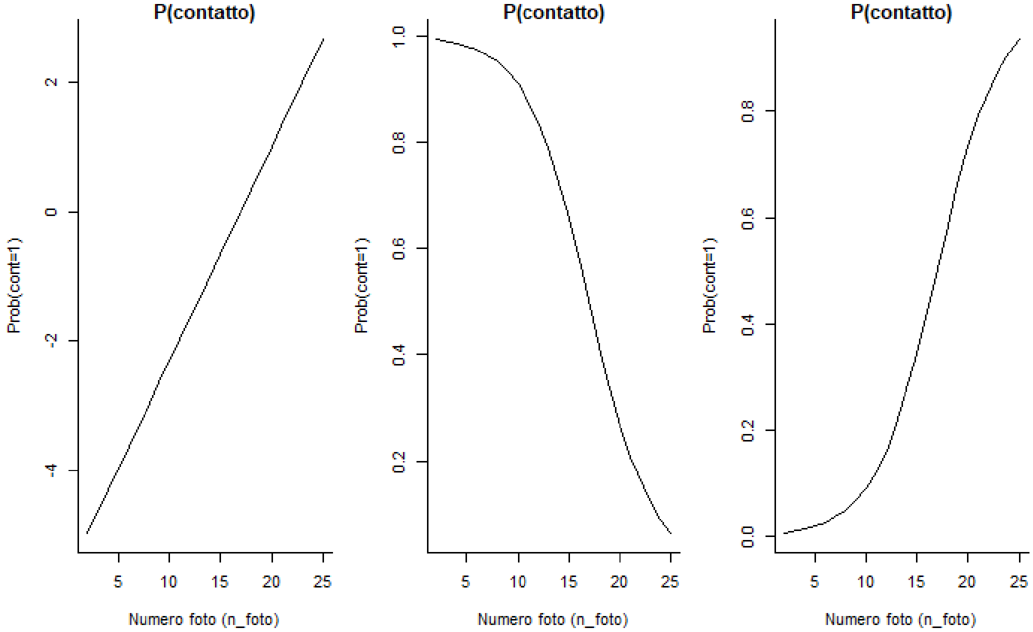
- `cont`: variabile dicotomica con valore 0/1. Il valore 1 indica che un possibile acquirente ha preso contatti con l'agenzia
- `n_foto`: il numero di foto inserite nel sito
- `m2`: la dimensione dell'immobile (in m\(^2\))
- `type`: l'informazione se l'immobile è una casa singola, una casa bifamiliare o un appartamento

La prima analisi si sofferma sull'effetto che ha il numero delle foto nell'annuncio (`n_foto`) sulla probabilità che un possibile acquirente abbia preso contatti con l'agenzia:

```
fit_foto <- glm(cont ~ n_foto, data = df,
                family = binomial)
summary(fit_foto)
```

```
Call:
glm(formula = cont ~ n_foto, family = binomial, data = df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.774  -0.371  -0.164   0.503   1.909

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.626      1.834   -3.07   0.0022 **
n_foto         0.332      0.103    3.22   0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.082  on 36  degrees of freedom
Residual deviance: 25.907  on 35  degrees of freedom
AIC: 29.91

Number of Fisher Scoring iterations: 6
```

La relazione stimata tra `n_foto` e `cont` viene mostrata in uno dei grafici sottostanti:




Viene poi derivato un test che verifica se $\beta_1$, il coefficiente del modello che descrive l'impatto del numero di foto sulla probabilità di un contatto, è uguale a 0.1:

$$H_0: \beta_1 = 0.1 \quad VS \quad H_1: \beta_1 \neq 0.1$$

Inoltre si ha:

```
confint.default(fit_foto,"n_foto")
```

```
         2.5 % 97.5 %
  n_foto 0.1297 0.5338
```

Infine, viene stimato un secondo modello in cui tutte le variabili presenti nel dataset sono incluse come variabili esplicative:

```
fit_all <- glm(cont ~ ., data = df,
               family = binomial)
```

e si desidera poi verificare quale modello fornisce una migliore bontà di adattamento:

```
logLik(fit_foto)
```

```
  'log Lik.' -12.95 (df=2)
```

```
logLik(fit_all)
```

```
  'log Lik.' -10.11 (df=5)
```

Si indichi quale grafico mostra la vera relazione tra `n_foto` e la stima della probabilità di osservare un incidente ottenuta nel modello `fit_foto`. Si indichi poi il valore della statistica test per la verifica di ipotesi indicata ($H_0: \beta_1=0.1$) e si indichi se l'ipotesi può o non può essere rifiutata. Si indichi se il valore dell'AIC per il modello `fit_all` risulta maggiore o minore di quello del modello `fit_foto`. Infine si indichi cosa è possibile riferire sulla relazione tra il tipo di casa nell'annuncio e la probabilità che vi sia un contatto da parte di un cliente.

a. ○ La stima ottenuto nel modello `fit_foto` è mostrata nel pannello A.

○ La stima ottenuto nel modello `fit_foto` è mostrata nel pannello B.

◉ La stima ottenuto nel modello `fit_foto` è mostrata nel pannello C. ✔

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: La stima ottenuto nel modello `fit_foto` è mostrata nel pannello C.

b. Il valore della statistica test per la verifica di ipotesi indicata: .

2,252

✔

c. ○ L'ipotesi non può essere rifiutata.

◉ L'ipotesi può essere rifiutata. ✔

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: L'ipotesi può essere rifiutata.

d. ○ AIC(fit_foto) > AIC(fit_all).

◉ AIC(fit_foto) < AIC(fit_all). ✔

○ AIC(fit_foto) = AIC(fit_all).

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: AIC(fit_foto) < AIC(fit_all).

e. ○ La probabilità di un contatto è più alta per le case bi-familiari.

◉ Non è possibile affermare quale tipo di immobile ha la probabilità più alta di contatto.. ✔

○ La probabilità di un contatto è più alta per le case signole.

○ La probabilità di un contatto è più alta per gli appartamenti.

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: Non è possibile affermare quale tipo di immobile ha la probabilità più alta di contatto..

La probabilità che avvenga un contatto deve per definizione avere un valore in (0,1) - questo esclude il pannello in cui la funzione ($P(contatto)$) ha valori al di fuori dell'intervallo (0,1). Il coefficiente che descrive l'effetto di `n_foto` su $P(Contatto)$ è positivo: questo indica che al crescere di `n_foto` cresce la probabilità di un contatto. Il livello di significatività usato è $\alpha = 0.05$.

La statistica test per un test quale $H_0: \beta_1 = \tilde{\beta} \quad VS \quad H_1: \beta_1 \neq \tilde{\beta}$ si calcola con $\frac{\hat{\beta}-\tilde{\beta}}{sd(\hat{\beta})} = \frac{0.3317-0.1}{0.1031}$ Il valore assoluta della statistica test va poi confrontato con il 97.5% quantile di una normale $z_{0.975} = 1.96$. In alternativa si può verificare se il valore 0.1 è contenuto nell'intervallo di confidenza (95%): se il valore è nell'intervallo l'ipotesi non può essere rifiutata.

$AIC(\mathcal{M}(p)): -2*logLik(\mathcal{M}(p)) + log * p$ quindi $AIC(\text{fit\_foto}) = 29.9071$ $AIC(\text{fit\_all}) = 30.223$ [Si potrebbe anche "semplicemente" controllare se la differenza tra la verosimiglianza di `fit_all` è maggiore o minore della differenza di gradi di libertà tra i due modelli (3).]

Non abbiamo elementi per giudicare quale tipo di casa ha un più alto tasso di contatto. L'informazione sarebbe deducibile avendo i valori dei coefficienti legati alla variabile `Type`, ma questi non sono disponibili.

a. Falso / Falso / Vero
b. 2.25
c. Falso / Vero
d. Falso / Vero / Falso
e. Falso / Vero / Falso / Falso

**Domanda 3**

Risposta corretta

Punteggio ottenuto 3,00 su 3,00

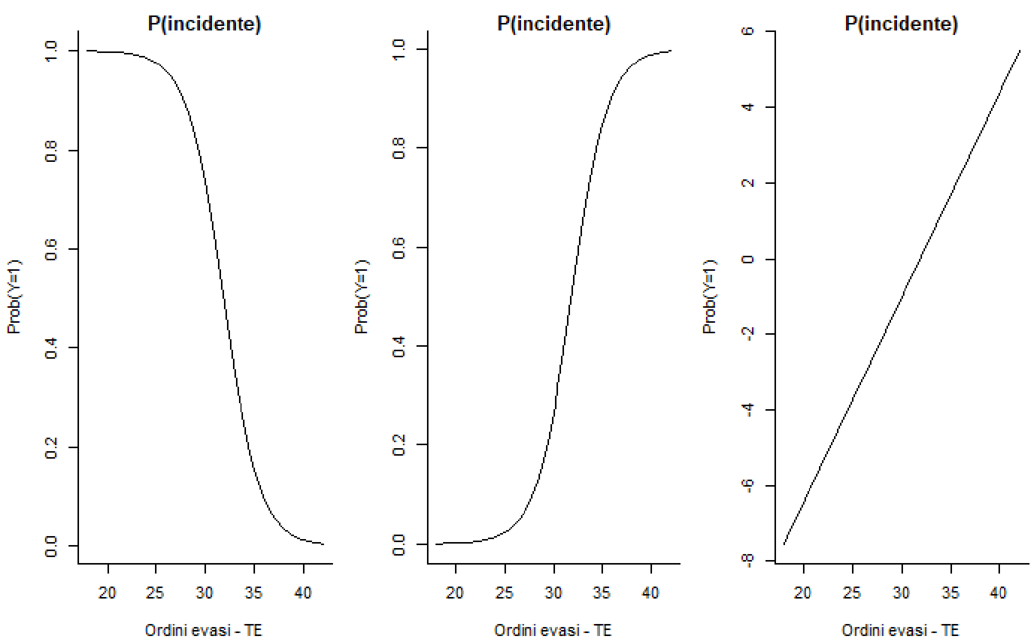
L'addetta alla sicurezza sul lavoro di una fabbrica tiene un record in cui indica se in una giornata sono registrati incidenti sul lavoro (una variabile dicotomica Y, in cui Y=1 indica che vi è stato un incidente). L'addetta desidera indagare se vi sono delle variabili esterne che influenzano la probabilità di registrare un incidente sul lavoro. Raccoglie quindi informazioni su potenziali fattori esterni quali il numero totale di ordini evasi nella giornata (TE) e alcune variabili climatiche (X1, X2, X3).

Il primo modello usato per indagare l'effetto del numero di ordini evasi sulla probabilità che avvenga almeno un incidente nell'impianto è un GLM (Binomiale) in cui solo la variabile \(te\) viene inserita nel predittore lineare:

```
fit1 <- glm(y~te, data = df, family = binomial)
summary(fit1)
```

```
Call:
glm(formula = y ~ te, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4611  -0.2544  -0.0674   0.4927   1.8866

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -17.296      4.170   -4.15  3.4e-05 ***
te             0.542      0.128    4.23  2.4e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99.591  on 71  degrees of freedom
Residual deviance: 44.202  on 70  degrees of freedom
AIC: 48.2

Number of Fisher Scoring iterations: 6
```

La relazione stimata TE e Y dal modello viene mostrata in uno dei grafici sottostanti:



Viene poi derivato l'intervallo di confidenza al 95% per il parametro $\beta_0$ (l'intercetta del modello):

```
confint.default(fit_te,"(Intercept)")
```

```
              2.5 % 97.5 %
  (Intercept) -25.47 -9.124
```

e si desidera poi verificare se un modello in cui vengono inserite anche le variabili climatiche fornisce una migliore bontà di adattamento:

```
fit_all <- glm(y~te+x1+x2+x3,family = binomial)
deviance(fit_te)
```

```
  [1] 44.2
```

```
deviance(fit_all)
```

```
[1] NA
```

Si indichi quale grafico mostra la vera relazione tra $te$ e la stima della probabilità di osservare un incidente ottenuta nel modello `fit_te`. Si indichi poi quale distribuzione viene usata per derivare l'intervallo di confidenza di $\beta_0$. Si indichi infine se il valore della devianza del modello `fit_all` è maggiore o minore del valore della devianza del modello `fit_te`.

a. La stima ottenuto nel modello `fit_te` è mostrata nel pannello A.
   La stima ottenuto nel modello `fit_te` è mostrata nel pannello B. ✔
   La stima ottenuto nel modello `fit_te` è mostrata nel pannello C.

   Punteggio ottenuto 1,00 su 1,00

   La risposta corretta è: La stima ottenuto nel modello `fit_te` è mostrata nel pannello B.

b. Per costruire intervalli di confidenza per $\beta_0$ si usa una distribuzione Chi-quadro.
   Per costruire intervalli di confidenza per $\beta_0$ si usa una distribuzione F.
   Per costruire intervalli di confidenza per $\beta_0$ si usa una distribuzione T-Student.
   Per costruire intervalli di confidenza per $\beta_0$ si usa una distribuzione Normale. ✔

   Punteggio ottenuto 1,00 su 1,00

   La risposta corretta è: Per costruire intervalli di confidenza per $\beta_0$ si usa una distribuzione Normale.

c. deviance(fit_all) < deviance(fit_te). ✔
   deviance(fit_all) > deviance(fit_te).

   Punteggio ottenuto 1,00 su 1,00

   La risposta corretta è: deviance(fit_all) < deviance(fit_te).

Dato che il coefficiente che descrive la relazione tra `te` e la probabilità di osservare un incidente è positivo si evince che al crescere di `te` aumenta la probabilità di osservare un incidente. Inoltre il grafico deve mostrare nella asse delle ordinate valori tra (0,1) e la tipica forma della curva logistica. La stima ottenuto nel modello `fit_te` è mostrata nel pannello B.

L'inferenza per i parametri nei GLM si basa sul fatto che le stime dei coefficienti di regressione sono stime di massima verosimiglianza e sono di conseguenza approssimativamente normalmente distribuite (per n $\rightarrow \infty$).

Più si aggiungono parametri al modello (cioè più si aumenta la complessità del modello) più il modello catturerà la variabilità dei dati (sebbene con il rischio che il miglioramento in termini di bontà di adattamento del modello sia minimo rispetto al costo di includere ulteriori parametri). Per questo motivo si ha che: deviance(fit_all) < deviance(fit_te).

a. Falso / Vero / Falso
b. Falso / Falso / Falso / Vero
c. Vero / Falso

**Domanda 4**

Risposta corretta

Punteggio ottenuto 5,00 su 5,00

L'addetto alla sicurezza sul lavoro di un gruppo industriale monitora il numero di incidenti avvenuti nelle 4 diverse fabbriche del gruppo. Per ogni fabbrica (variabile FAB) è disponibile il numero di incidenti settimanali (Y) e il numero totale di ordini evasi nella settimana in decine (TE).

```
summary(df)
```

```
       y             fab          te
 Min.   : 2.00   Bari  :15   Min.   :20.3
 1st Qu.: 5.00   Nola  :15   1st Qu.:33.1
 Median : 7.00   Rho   :15   Median :43.5
 Mean   : 8.52   Torino:15   Mean   :42.1
 3rd Qu.:11.00               3rd Qu.:49.9
 Max.   :26.00               Max.   :58.6
```

Si desidera indagare se il numero di ordini evasi ha un effetto sul numero di incidenti e l'importanza di questo effetto nelle diverse fabbriche. Il primo modello preso in esame è `fit_add`, un GLM (Poisson) in cui il predittore lineare include l'effetto di \(te\) e \(fab\) in maniera additiva:

```
fit_add <- glm(y~te+fab, family = poisson)
summary(fit_add)$coef
```

```
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)  0.37323   0.219033  1.7040 8.838e-02
te           0.03931   0.004475  8.7838 1.581e-18
fabNola      0.39642   0.119357  3.3213 8.960e-04
fabRho      -0.30071   0.136824 -2.1978 2.796e-02
fabTorino   -0.07950   0.132849 -0.5985 5.495e-01
```

```
fit_add$df.residual
```

```
[1] 55
```

Usando questo modello si desidera capire quale fabbrica ha il più alto numero di incidenti a parità di ordini evasi.

Si desidera poi valutare se vi sia evidenza che l'effetto della variabile \(te\) sia diverso da fabbrica a fabbrica. Per valutare l'ipotesi viene stimato il modello `fit_mult` e viene poi applicato un LRT, il cui p-value è riportato:

```
fit_mult <- glm(y~te*fab, family = poisson)
anova(fit_add, fit_mult, test = "LRT")$`Pr(>Chi)`[2]
```

```
[1] 0.2874
```

L'addetto desidera infine stimare, usando il modello `fit_add`, il numero di incidenti in una settimana in cui \(te=40\) nelle fabbriche di Bari e Torino:

```
fit_add$coef
```

```
(Intercept)          te     fabNola      fabRho   fabTorino
    0.37323     0.03931     0.39642    -0.30071    -0.07950
```

```
predict(fit_add, data.frame(nd = "Bari", te = 40),type = "response")
```

```
[1] NA
```

```
predict(fit_add, data.frame(nd = "Torino", te = 40),type = "response")
```

```
[1] NA
```

Si indichi quale fabbrica ha il più alto numero di incidenti a parità di ordini evasi. Si indichi quale interpretazione dare al risultato del LRT. Si indichino i valori stimati del numero di incidenti nelle fabbriche di Bari e Torino (cioè i valori mancanti delle funzioni `predict`). Infine si indichi il numero di gradi di libertà residui del modello `fit_mult` (cioè l'output di `fit_mult$df.residual`).

Il livello di significatività usato per i test è \(\alpha = 0.05\).

a. ○ A parità di ordini evasi, Bari ha in media il maggior numero di incidenti.

◉ A parità di ordini evasi, Nola ha in media il maggior numero di incidenti. ✔

○ A parità di ordini evasi, Rho ha in media il maggior numero di incidenti.

○ A parità di ordini evasi, Torino ha in media il maggior numero di incidenti.

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: A parità di ordini evasi, Nola ha in media il maggior numero di incidenti.

b. ○ Vi è evidenza che l'effetto di TE sul numero di incidenti sia diverso nelle fabbriche.

◉ Non si può affermare che ci sia evidenza che l'effetto di TE sul numero di incidenti sia diverso nelle fabbriche . ✔

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: Non si può affermare che ci sia evidenza che l'effetto di TE sul numero di incidenti sia diverso nelle fabbriche .

c. Il valore stimato di incidenti a Bari usando il modello `fit_add` quando te=40.

6,998

✔

d. Il valore stimato di incidenti a Torino usando il modello `fit_add` quando te=40.

6,463

✔

e. Il numero di gradi di libertà residui per `fit_mult`.

52

✔

A parità di ordini evasi la fabbrica con il valore di intercetta più alto ha il maggior numero di incidenti. Il valore dell'intercetta per la fabbrica di Bari corrisponde al valore dell'intercetta del modello, mentre per le altre fabbriche è necessario sommare il valore del coefficiente per i singoli livelli visibile nell'output di summary. Valori negativi del coefficiente indicano fabbriche in cui a parità del valore di `te` ci sono meno incidenti rispetto a Bari. Valori positivi del coefficiente indicano fabbriche in cui a parità del valore di `te` ci sono più incidenti rispetto a Bari.

La differenza tra il modello `fit_add` e `fit_mult` è che in `fit_mult` l'effetto di `te` sul numero di incidenti può essere diverso per ogni fabbrica: un valore alto del LRT (cioè un valore basso del pvalue) indicherebbe una forte evidenza che l'effetto di `te` sia effettivamente diverso nelle varie fabbriche. Dato che il modello `fit_mult` aggiunge tre possibili diversi coefficienti che descrivono l'effetto di `te` sugli incidenti nelle fabbriche che non sono Bari il numero di gradi di libertà residui del modello è: `fit_add$df.residual` - 3, 52.

Il modello usa il coefficiente canonico (cioè il default), che per la distribuzione di Poisson è il logaritmo. Si ha quindi:

- il valore stimato a Bari: \[\exp\{0.3732 + 1.5723\}\]
- il valore stimato a Torino: \[\exp\{0.3732 + (-0.0795) + 1.5723\}\]

a. Falso / Vero / Falso / Falso
b. Falso / Vero

c. 7.00
d. 6.46
e. 52.00

◄ Quiz MLR

Vai a...

Esercizi modelli lineari ►